REPORT NO. 2

31 JANUARY 1963

RESEARCH IN INFORMATION RETRIEVAL

Second Quarterly Report

1 October 1962 - 31 December 1962

Contract No. DA 36-039-SC-90787
File No. 1160-PM-62-93-93(6509)

Technical Report P-AA-TR-(0031)

U. S. Army Electronics Research and Development Laboratory
Fort Monmouth, New Jersey

ITT International Electric Corporation
Route 17 and Garden State Parkway, Paramus, New Jersey
A SUBSIDIARY OF INTERNATIONAL TELEPHONE AND TELEGRAPH CORPORATION

Report No. 2 31 January 1963

RESEARCH IN INFORMATION RETRIEVAL

Second Quarterly Report
1 October 1962 - 31 December 1962

An investigation of the techniques and concepts of information retrieval

Contract No. DA 36-039-SC-90787
File No. 1160-PM-62-93-93(6509)

Signal Corps Technical Requirement
SCL-4218 12 January 1960

Technical Report P-AA-TR-(0031)

Jacques Harlow, Principal Investigator

prepared by
Alfred Trachtenberg
Quentin A. Darmstadt
George Greenberg

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

1. PURPOSE

1.1 SCOPE

This report discusses the work performed for the U. S. Army Electronics Research and Development Laboratory under Contract No. DA 36-039-SC-90787 during the period from 1 October 1962 to 31 December 1962.

1.2 OBJECTIVES

The objective of this project is to investigate the techniques and concepts of information retrieval and to formulate and develop a general theory of information retrieval. The formalization of this theory is oriented to the automation of large-capacity information storage and retrieval systems. This theoretical framework will be the basis for the utilization of general purpose stored-program digital computer systems for performing the storage and retrieval functions.

1.3 PROJECT TASKS

During the first quarter of this project a preliminary model of the information storage and retrieval problem was developed as a frame of reference for subsequent analysis. This quarter was spent in more detailed investigations of significant aspects of the problem as related to the transformational functions of the model.

In the analysis of any complex problem there are essentially three levels of understanding to master: the whole, the parts, and the relation of the parts to the whole. The preliminary model constitutes the whole; the transformation functions comprise the parts. However, there are a number of alternative approaches that may be considered for each part or function. These approaches become the specific tasks or subtasks of the project.

At this stage many ramifications of the transformational functions have been analyzed. Although these studies pertain to manifest tasks, they have not been formally designated as such. The process of formalization depends upon a review of the relation of each part to the central problems of the whole. Specific tasks will be assigned during the next quarter, and subsequent reports will be oriented to the activity performed under these designated tasks.

This discussion does not vitiate the statement in this section of the First Quarterly Report. In that report three tasks were defined; but these tasks pertain to methodology rather than functional requirements. At this stage of the project it is essential to shift from a methodological to a functional viewpoint.

2. ABSTRACT

This report discusses research activity performed in the investigation of the techniques and concepts of information retrieval. The general problems of information storage and retrieval are reviewed to establish a framework for the development of general theoretical principles. Several functional characteristics of the preliminary model--the representation of file items, file organization, system design and synthesis, and relevance--are summarized in terms of tentative solutions and their attendant difficulties. Specific aspects of the problem--information theoretical methods of document categorization and corrective procedures for automatic indexing--are examined in detail.

3. PUBLICATIONS, REPORTS, AND CONFERENCES

3.1 TECHNICAL NOTES

The following internal technical memoranda were issued during this reporting period:

(a) IEC TECHNICAL NOTE, File No. P-AA-TN-(0043)-N, 18 December 1962; A Measure of Effectiveness for Document Retrieval Systems, Quentin A. Darmstadt.

(b) IEC TECHNICAL NOTE, File No. P-AA-TN-(0044)-N, 20 December 1962; Corrective Procedures for Automatic Indexing Systems, Alexander Szejman.

(c) IEC TECHNICAL NOTE, File No. P-AA-TN-(0045)-N, 27 December 1962; Information Theoretical Methods of Document Categorization, Alfred Trachtenberg.

(d) IEC TECHNICAL NOTE, File No. P-AA-TN-(0046)-N, 31 December 1962; Survey of Mathematical Models of Various Aspects of Information Retrieval, Quentin A. Darmstadt.

These technical notes are dated at the time of their completion; these dates do not necessarily correspond to the date of publication.

3.2 REPORTS

The following reports were issued during this reporting period:

(a) RESEARCH IN INFORMATION RETRIEVAL: First Quarterly Report, 1 July 1962 - 30 September 1962, Technical Report P-AA-TR-(0010), (Manuscript Version), 30 October 1962.

(b) MONTHLY LETTER REPORT NO. 3, 1 October 1962 - 31 October 1962, File No. P-AA-TR-(0012), 31 October 1962; Research in Information Retrieval, Alfred Trachtenberg.

(c) MONTHLY LETTER REPORT NO. 4, 1 November 1962 - 30 November 1962, File No. P-AA-TR-(0025), 30 November 1962; Research in Information Retrieval, Alfred Trachtenberg.

3.3 CONFERENCES

The following conferences were held between IEC and USAELRDL personnel:

(a) 29 November 1962--Meeting at IEC. IEC personnel were introduced to Mr. Anthony V. Campi, who had recently been assigned as Project Engineer. Several aspects of the First Quarterly Report were discussed, and the concepts pertaining to measure of relevance were clarified. IEC accepted the suggestion that the discussion in the report should be elaborated in more detail.

Mr. Quentin A. Darmstadt attended the conference entitled "Mathematics of Information Storage and Retrieval," which was conducted by Dr. Robert M. Hayes under the auspices of the Georgia Institute of Technology from 3 to 7 December 1962. The relevance of the conference to this project is evident in the title. However, because of general significance of the conference, attendance was sponsored by IEC.

4. FACTUAL DATA

4.1 STATEMENT OF THE PROBLEM

The technical requirement for this project, as stated in SCL-4355, specifies "...a research investigation of techniques and concepts necessary for the efficient mechanization of large-capacity information storage and retrieval systems." The future applied objectives suggested as guides for such research constitute a range of "...problems of military significance; i.e., personnel files, intelligence data, etc."

The problem as presently conceived is to develop a general theory of information retrieval whose primary goal is its use as a system tool for the optimum design of specific information retrieval systems in the future. The project is oriented to a theory of systems that can be applied to the design of specific job oriented systems in their entirety rather than to a specific procedure(s); to dealing with real contexts that may be of interest to the Army, wherever possible, rather than necessarily limiting the study to abstract formalism; to the consideration of optimum hardware once software at the level of algorithm rather than machine code has been specified; and to the problem of conversion to canonic form when linguistic complexity is not the critical problem.

A general model of the information retrieval process has been developed. This model provides a framework both for understanding the critical features of information retrieval systems of different levels of sophistication and for isolating critical areas of information retrieval procedures and techniques to focus upon for further development.

4.2 SYSTEM MODEL

4.2.1 Analytic Framework - The information retrieval model developed in the First Quarterly Report forms the basis for the analytic framework. This model defines information storage and retrieval formally and abstractly, although the model is quite simple. The three algorithmic transformations isolated (D, E, P) do not presuppose any specific form of classificatory or interrogatory vocabulary, nor do they depend upon any unique search procedures or file structure. Furthermore, there is no precommitment in allocating the functions to manual or machine processing.

Ultimately, the model should encompass a completely automated information content storage and retrieval system. Such a system is infeasible in the present state-of-the-art of automating human cognitive functions. Only the processing or P transform for limited document retrieval, with fairly imprecise but humanly generated indices, is currently being automated. Even for this limited application the logical file organization and search procedures as well as their implementation can be substantially improved.

The study of sophisticated file organization and search procedures for traditional information retrieval systems will continue to be an aspect of this program. Even more important, however, will be the development of file organizations and search procedures for the efficient implementation of system capabilities that will have to evolve before fully automated information content storage and retrieval systems can be developed.

These new capabilities include the automation of functions that can currently be performed only by people and the development of explicit transformation algorithms for the model. One of the most difficult areas for automation is the formalization of ordinary language to describe the information in a form suitable for efficient storage and effective retrieval. This problem pertains to the input and query, or D and E transforms, respectively. The question of linguistic analysis per se has been deemphasized. However, the more general problem of improving and automating the D and E transforms is essential to the goals of the project.

There are a number of relatively discrete capabilities that will have to be developed, primarily in the input and query transforms. It is possible to describe several procedurally oriented tasks for producing these capabilities. Each of the model transforms, D (data input), E (query), P (processing), and D^{-1} (output), will be considered in turn.

4.2.2 The D Transform - The central problem in the transformation of information inputs to forms usable in storage and retrieval is one of classifying, categorizing, or indexing. To date all operational classificatory schemes tend to be intuitively formulated, manually implemented, and statically evolved; these schemes are virtually impossible to change systematically.

There are, therefore, three areas in which further capabilities must be developed:

(a) Explicit procedures for establishing useful category groupings and boundaries.

(b) Definitive procedures for automatically assigning items or documents to index categories accurately and efficiently.

(c) Methods for improving the precision of indexing.

The methods for improving precision include adaptive procedures for altering index assignments to align document categories more closely with the users' categories as a function of feedback on the adequacy of individual searches.

These capabilities are in some measure mutually interdependent and cannot ultimately be developed without reference to other system transforms. Similarly, the capabilities of other system transforms will impinge upon the organization of the D transform. Thus the development of useful and efficient category groupings of descriptors or indices may be best considered in relation to specific schemes for automatic document classification. The work of Borko and Bernick [6] illustrates this approach. Similarly, the validity of adaptive procedures for reorganizing descriptor assignment is clearly dependent upon the techniques, automatic or manual, used to assign item categories initially.

It is important to note, however, that these three capabilities are distinct; work may proceed relatively independently with reasonable expectation of later integration into a system concept. The work of Borko and Bernick fails to demonstrate that joint consideration of automatic category generation and automatic category assignment results in either an improved category structure or an improved prediction scheme. Furthermore, attacking these problems as separate capabilities may be advantageous in allocating effort more efficiently and in developing more general techniques. Thus work on the problem of finding ideal categories for grouping items into larger categories may result in techniques for decompassing larger items into coherent smaller units or categories. The latter problem is part of the more general problem of developing explicit procedures for establishing useful categories and their boundaries--regardless of the level of organization between or within items that the categories refer to.

This discussion does not imply that work on the explicit generation of useful categories should necessarily be unconcerned with adequate automatic prediction of *a priori* categories. The significant point is that each task should focus upon the development of as powerful a capability as possible. If work in one area suggests an approach to any other, then so much the better.

The formal development of each of these problems is continuing. Various techniques such as the theory of clumps [21], factor analysis [5], and latent class analysis [1] have been suggested for dealing with automatic category generation. These techniques are being evaluated together with the concepts presented in subsequent sections. The evaluation is essential for the ultimate selection of the most useful procedure for categorization.

4.2.3 The E Transform - The E transform is the set of algorithms that transposes the users' queries to the processor. In an ideal system the E transform would handle any query couched in the natural language of the user. The present state-of-the-art in information retrieval is far too primitive to deal with any sophisticated query. Except for specialized files such as those developed for Baseball [11], ACSIMATIC [23], or the multi-list system of Prywes and Gray [9, 10], questions of fact cannot be answered by contemporary information storage and retrieval systems.

Both Baseball and ACSIMATIC do contribute to the conceptual basis of the E transform. Baseball analyzes English query sentences, and ACSIMATIC provides a uniquely articulated query format appropriate to the intelligence problem. However, both are inappropriate to the general information storage and retrieval problem in their present status. The contributions of Prywes and Gray are not pertinent because the problem they address is primarily in the area of file organization for attribute-value data. While Prywes and Gray do not contribute to the problem of the E transform, their work is important relative to the P transform.

These statements are not intended to be derogatory nor to denigrate the significance of these projects. Therefore, further clarification is warranted. There are two kinds of fact retrieval:

(a) The retrieval of facts from a table or file specifically organized by the inventiveness of human programmers for the retrieval of the summarized facts.

(b) The retrieval of facts or content, the implicit goal of the preliminary model, from items or documents couched in ordinary language.

The three cited systems all deal with restricted and specifically organized data--baseball scores, combat intelligence, and personnel files. One approach to the direct retrieval of facts from documentary items is to assume that the problems of the P and E transforms, as specified for these systems, are essentially solved. Then the only remaining difficulty is to reduce facts in ordinary language to the proper tabular or list form.

To adopt this approach, however, is counter-evolutionary. The burden of development remains in the area of the D transform. In order to transform informal data automatically into the format required by these systems, an inordinately long time may pass without any significant advances toward the goal of automated content retrieval.

At present the only query allowed for documentary data is: "What documents contain information of the following kind: ________?" This limitation on queries has many shortcomings. Not all of these shortcomings must be overcome simultaneously; an evolutionary approach would focus upon expanding query capability by isolating specific problem areas and concentrating on them.

There are several important shortcomings or, conversely, desirable capabilities. The first is a limitation to documents. The query capability should be extended so that a system could respond with either large bounded portions of larger documents or with an automatically generated extract or abstract of the relevant facts in the document. As these capabilities are developed, a system will approach the goal of allowing questions of the form: "What information do you have on...;" rather than: "What documents...."

The second shortcoming pertains to a limitation to all documents containing relevant information. It is practical not to retrieve information from or about all documents. If a large number of documents cover a narrow specialized subject, the relevant information may be scanty, redundant, or qualitatively poor. In such cases it would be beneficial to restrict the scope of retrieval or, initially, indexing.

Finally, there is a limitation, in the extreme, on the characterization of the information intended by the conditional phrase, "...of the following kind: ________." Different operational information systems impose different limitations of this type. A hierarchically organized index or query language may produce such unusual classifications of new material that a subsidiary index is necessary in order to use the primary index properly. Freer Uniterm systems are limited to Boolean functions of two-valued descriptors; the descriptor is either present or absent. The use of role indicators [22] and similar devices [31] offer some possibility of improving the query. But the crux of the problem is to develop a query capability that allows a user to state his question precisely. This ability is essential to useful content retrieval.

The three problem areas cannot produce a content retrieval system if attention is restricted to the E transform. The P transform must evolve to be able to handle more sophisticated queries. Similarly, the organization of the D transform must be capable of generating the required categories and preserving the information for a range of anticipated queries. Thus work on the categorization aspect of D transform is critical if items of smaller scope than an entire document are to be automatically isolated. Similarly, the methods for improving indexing are essential to improving the precision of the users' queries.

The interdependence between the D and P transforms does not invalidate approaching the problems of the E transform. The major problem with some of the more sophisticated information systems is that so little thought was given to the query process. The result is systems that are too cumbersome to use. It is essential that the intentions, requirements, and capabilities of potential human users be carefully analyzed before the organization of D and P transforms for future systems are fully established. For some kinds of information retrieval such as general education and scholarly research the open stacks and card catalogues of present libraries suffice. For other information retrieval problems such as keeping abreast of new developments or resolving specific matters of fact, innovations are vitally necessary. But such innovations are valueless unless the system allows the user to ask intelligent and appropriate questions.

4.2.4 The P Transform - Advances in information storage and retrieval depend upon improved processing algorithms. Unfortunately advances in the other transforms will influence the choice of processing techniques. It is, consequently, difficult to define relatively independent problem areas.

A basic study continues in the analysis of processing requirements for traditional systems and for new capabilities as they become evident. Among the subjects that have been analyzed relative to the P transform are:

(a) Measures of relevance and their processing applications.

(b) Measures of efficiency and their optimization.

(c) Measures of cost for both successes and failures.

(d) Search theory and procedures.

(e) File structure and organization.

(f) System synthesis.

Obviously, this list is heterogeneous and requires further elaboration and refinement. Some subjects are intimately related to other system transforms and thus depend upon the outcome of advances in these transforms. Others are supraordinate in nature and are, therefore, perhaps best deferred until a specific system has been designed. A general approach to these subjects may be possible; since such an approach would have the greatest impact on the processing configuration, these subjects were included as tentative functions of the P transform.

4.2.5 The D^{-1} Transform - No substantive elements of this transform have been defined.

4.3 INFORMATION THEORETICAL METHODS OF DOCUMENT CATEGORIZATION

4.3.1 General - This section presents some applications of information theory to the problem of document classification or categorization. Criteria for a good categorizer are presented, and various information theoretical measures that measure the goodness of categorizers are examined.

The problem of document categorization is the problem of selecting from a set of possible categories those categories to which a document may belong. This selection would have to be based upon certain clues or indications found in the document itself. Thus, as Maron [17] has stated, the problem of categorization can be divided into two parts: the selection of certain relevant aspects of a document as clues toward classification; and the use of these clues to predict the proper category to which the document belongs. Once the method of classification has been defined, then the procedures could be automated.

Many authors [1, 2, 5, 7, 16, 20, 25] have felt that the occurrence of certain words in a document provided excellent indications of the category to which that document belonged. Based upon word occurrence statistics, document categories would be predicted automatically. This approach is also developed here, but certain information theoretical techniques are applied that do not appear to have been applied elsewhere.

This approach assumes that a group of human experts will initially classify a number of documents into a given set of categories. A basic assumption is that all categories that receive one or more documents will be retained as permanent categories, which will be the only categories used in the future. Another assumption is that the number of documents initially classified by experts is large enough so that the statistics of this group may be assumed to reflect the statistics of the body of documents that may later be automatically categorized. In other words, relative frequencies of categorization obtained from the initial group will be used as the probabilities of categorization of the larger group.

4.3.2 Criteria for Selecting Predictors - It is expected that the occurrence of certain words in a document indicates the categorization of that document. It follows that one of the criteria for selecting a particular word to predict categories is that its occurrence in documents be strongly correlated with the appearance of those documents in a particular category--for those documents that were initially classified. In other words, a word that appears in every document of a particular category and appears in no document of any other category seems to be an ideal predictor of that category. In practice there may be few of these ideal predictors; then it is necessary to look for words for which occurrence in a document means a particular category for that document is much more likely than any other category.

This criterion would be sufficient for choosing indicator words if the distribution of documents in the categories were uniform. In practice, this condition would generally not be the case; some categories would have many more documents than others. Then a word that would seem to be an excellent indicator might be found to supply no more information than the total distribution of documents supplied. Thus the occurrence of the good indicator word in documents must not only be strongly correlated with the classification of these documents in one particular category, but the distribution of documents containing this word must also markedly differ from the distribution of all the documents.

4.3.3 Information Theoretical Treatment of Predictor Criteria

4.3.3.1 Statement of the Problem - The problem can now be expressed mathematically: Given N documents* classified into C_j categories, where $j = 1,...k$. The vocabulary of the N documents contains m words, W_i, $i = 1,...m$. Word W_i occurs in N_i documents, and n_{ij} of these

*The classification of a document into two or more categories is counted as the classification into one category each of two or more documents.

documents fall into category C_j.

Let:

$p(C_j)$ = the probability that a document falls into category C_j

$p(C_j|W_i)$ = the probability that a document with the word W_i falls into category C_j.

Then: $$p(C_j) = p_j = n_j/N \qquad (4\text{-}1)$$

and: $$p(C_j|W_i) = p_{ij} = n_{ij}/N_i \qquad (4\text{-}2)$$

The following relationships hold by definition:

$$\left.\begin{array}{l} \sum\limits_j n_{ij} = N_i \\ \sum\limits_j n_j = N \\ \sum\limits_j p_j = \sum\limits_j p_{ij} = 1 \end{array}\right\} \qquad (4\text{-}3)$$

It has been assumed that there exists at least one document in each category; i.e., the smallest possible $p_j = 1/N$. If there were no documents in a category C_e, then p_e would be zero; consequently, all the p_{ie} would be zero. Such a category would be of no use and would be discarded. Having at least one document in each category also implies that $k \leq N$, and that the largest possible $p_j = 1 - \frac{k-1}{N}$; for there are k - 1 categories that would have to have the minimum p_j. Therefore:

$$\left.\begin{array}{ll} & \frac{1}{N} \leq p_j \leq 1 - \frac{k-1}{N} \\ \text{and:} & 0 \leq p_{ij} \leq 1 \end{array}\right\} \qquad (4\text{-}4)$$

4.3.3.2 <u>Definitions of Measures of Goodness</u> - The <u>non-correlation</u>

of word occurrence and category or the uncertainty of category, given the occurrence of a word W_i, can be expressed by Shannon's formula for entropy:

$$H_i = H(C_j|W_i) = - \sum_j p_{ij} \log p_{ij} \qquad (4\text{-}5)$$

Thus a good indicator word would have a low H_i. But is this word supplying more information than the total document distribution? Maron suggest a measure:

$$M_i = H - H_i \qquad (4\text{-}6)$$

where: $$H = H(C_j) = - \sum_j p_j \log p_j \qquad (4\text{-}7)$$

H is simply the uncertainty of categorization when no word occurrences are known; that is, H is the entropy of the a priori distribution of all of the documents.

This measure, however, does not seem adequate. Difficulty arises when the a priori p_j are unequal and have the same numerical value as the p_{ij} of different categories; in this case, $H = H_i$ and $M_i = 0$, which indicates a bad predictor; but W_i may actually be a good one in terms of the given criteria. The example in Figure 1 illustrates this difficulty. Clearly $H = H_r$ and $M_1 = 0$ in Figure 1, but W_r is a good predictor and supplies a great deal of information.

More effective measures of the adequacy of an indicator word can be based on a relative entropy function of the type found in Watanabe [32]. This function is similar to the previous entropy functions, but it accounts for the a priori probabilities directly. The relative entropy, S_i, is defined by:

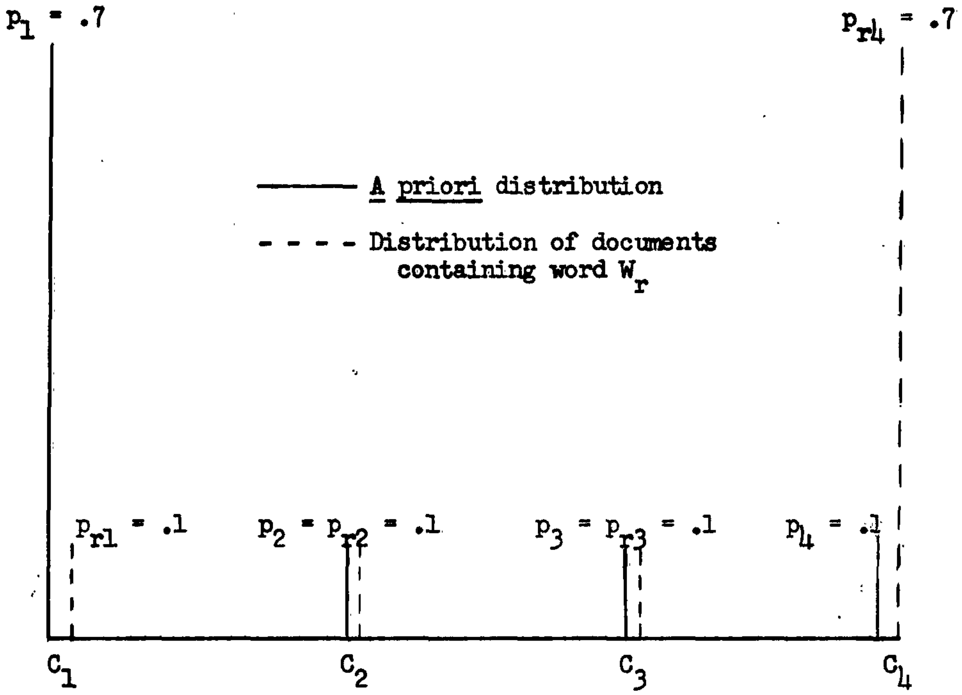

FIGURE 1. Probability Distributions for a Class of Documents

$$S_i = S(C_j|W_i) = -\sum_j p_{ij} \log \frac{p_{ij}}{Ap_j} \qquad (4\text{-}8)$$

where A is a positive constant chosen to keep S_i non-negative. A should be chosen such that $A = 1/p_e$, where $p_e \leq p_j$ for all j, so that $S_{imin} = 0$. This condition means that $k \leq A \leq N$, since $1/N \leq p_e \leq 1/k$.

Before these measures are defined and examined, one more entropy function must be defined:

$$H_A = -\sum_j p_j \log p_j/A = H + \log A \qquad (4\text{-}9)$$

Three possible measures will now be defined, in addition to the measure M_1 that Maron has suggested.

$$\left.\begin{aligned} M_1 &= H - H_i \qquad \text{(Maron's measure)} \\ M_2 &= H - S_i \\ M_3 &= H_A - S_i \\ M_4 &= \log A - S_i \end{aligned}\right\} \qquad (4\text{-}10)$$

Now:

$$\left.\begin{aligned} M_2 &= H - H_i - \sum_j p_{ij} \log p_j - \log A \\ M_3 &= H - H_i - \sum_j p_{ij} \log p_j = M_2 + \log A \\ M_4 &= - \sum_j p_{ij} \log p_j - H_i = \sum_j p_{ij} \log \frac{p_{ij}}{p_j} \end{aligned}\right\} \qquad (4\text{-}11)$$

The new M_2 and M_3 are similar to M_1, except for a cross-term that relates the p_j and the p_{ij}. M_4 also has this cross-term. M_3 is simply M_2 with the constant term missing.

4.3.3.3 Maxima and Minima of the Measures of Goodness - The behavior of these measures of goodness and the various entropy functions are developed in Appendix A, Section 8.

4.3.3.4 Evaluation of the Measures - Measure M_1 was shown to be inadequate, since it may erroneously indicate that a good predictor is a bad predictor. In addition, M_1 can assume negative values. M_2 can also assume negative values, which may make it inconvenient to use. M_2 is also inconvenient to calculate, since it requires the calculation of two sums, $\sum_j p_j \log p_j$ and $\sum_j p_{ij} \log \frac{p_{ij}}{p_j}$, and since the last summation also includes a division operation. M_3 requires the calculation of these same sums, although it is slightly more convenient to use since M_3 is always positive. M_1, M_2, and M_3 have fairly complex expressions for maxima and minima; M_1 and M_2 become negative and M_3 never reaches zero.

It seems clear then that M_4 is the best measure of the group: it is always positive, has a simple expression for a maximum, has a zero minimum, and is easier to calculate than the others.

4.3.3.5 Mathematical Expression of Predictor Criteria - The correlation of the occurrence of an indicator word in a document and the classification of that document in a particular category would be measured by H_i.

$$H_i = - \sum_j p_{ij} \log p_{ij} \qquad (0 \leq H_i \leq \log k) \qquad (4\text{-}12)$$

A low H_i indicates a good predictor; a high H_i, a bad predictor.

A measure that also accounts for the a priori distribution of documents and indicates how much more information the predictor supplies than this distribution is M_4.

$$M_4 = \sum_j p_{ij} \log \frac{p_{ij}}{p_j} \qquad (0 \leq M_4 \leq - \log p_e) \qquad (4\text{-}13)$$

$$(1/N \leq p_e \leq 1/k)$$

A high M_4 indicates a good predictor; a low M_4, a bad one. Both of these measures must be taken into account when choosing indicator words.

4.3.4 Predictors - On the basis of these mathematical criteria, it is now possible to select clues or predictors. A word that has a high value for M_4 and a low value for H_i will be selected. The cutoff point for these functions for good predictors must be determined experimentally. It is difficult to say how high a value for M_4 or how low a value for H_i is actually needed for a good predictor without empirical verification.

Not only can single words be used as predictors, but word pairs, word triplets, and higher word combinations can also be used with an expected improvement in prediction. The mathematics for these cases is essentially the same; the only difference is that the occurrence of word pair $[W_a\ W_b]$ or word triplet $[W_a\ W_b\ W_c]$ is considered instead of the single word W_i. These word pairs and word triplets can be ranked together with single words on the same scale, and their effectiveness as predictors can then be compared.

4.3.5 Application of Clues to Predicting Categories - Once the significant predictors have been determined, it is possible to obtain the probability that a document appears in a category on the basis of those predictors. This probability is:

$$p(C_j | W_a\ W_b \ldots\ldots) \qquad (4\text{-}14)$$

Maron gives an approximation to this probability. In general, this approximation would require a great deal of calculation. One way of approximating the probability would be to take the weighted average of the category probabilities using each of the most significant indicator words. Other functions of these words might also approximate the probability. Thus, in general, the predicted category would be some function of the category probabilities for each of the words. Methods for determining suitable functions of this kind should be investigated.

4.3.6 Modification of Categories - Implied in this discussion are criteria for modifying and combining categories to get better classification. What is needed is a set of categories that would be strongly correlated with word occurrence and that would yield approximately equal *a priori* category probabilities. In this way, there would be words with high M_4 and low H_i. In fact, these two measures would then be almost the same; for if $p_j \doteq 1/k$ for all j, then:

$$M_4 \doteq \sum_j p_{ij} \log p_{ij} + \log k = \log k - H_i \qquad (4\text{-}15)$$

Thus in equalizing the categories, if for some W_i, M_4 is high and there exists at least one such W_i for each category, then the classification would be a good one.

4.3.7 Summary - The criteria for selecting appropriate words in a document as predictors of the document category have been presented. Representations of these criteria have been demonstrated in terms of information theoretical measures. These measures have been analyzed and evaluated; one set designated as M_4 and H_i was finally chosen as the most effective. An indication of how the category might be selected was then developed; similarly, an indication of the basis on which the existing categories might be modified to improve classification was suggested.

Although this discussion has been presented in terms of selecting one of k major categories, once a major category has been determined, the same process can be used to determine subcategories; the mathematics are identical, and subcategory statistics would be used.

4.4 CORRECTIVE PROCEDURES FOR INDEXING SYSTEMS

4.4.1 General - This section investigates the methods and feasibility of applying corrective procedures to indexing systems. A fundamental aspect of these concepts is their ultimate adaptability to automated procedures. The first part of this discussion presents the basic ideas of this concept; the second part develops the concept formally.

4.4.2 The Taxonomy of Indexing Systems - Information retrieval systems consist of a library of documents and set of indexing rules and procedures for linking descriptors to documents. The documents in this context refer to the smallest ensemble of information subject to retrieval; these documents are considered as being indivisible. The indexing rules and procedures theoretically select descriptors that bear some relation to the descriptors used by people who will interrogate the system.

The system may accept new documents in its library; the documents are then classified according to the rules and procedures of the indexing scheme of the system. The system is not necessarily committed to the use of old descriptors. The indexing rules allow for the supply of new descriptors with the acceptance of the new documents by the library.

The user specifies his requests for information by writing a sequence of acceptable descriptors in the form of a Boolean function; that is, the descriptors are joined by OR and AND. The user's disposition of the descriptors implies the existence of an ideal taxonomic system. The taxonomy imposed by the indexing rules and procedures constitutes an external taxonomy or a priori taxonomy.

A corrective procedure will cause the external taxonomy to evolve into the ideal taxonomy on the basis of information concerning the adequacy of the sets of documents retrieved. This information is supplied by the user.

The central problem is: On what factors does the functioning of a corrective procedure depend? The answer to this problem depends upon the elucidation of the relation between the ideal and the external taxonomy. More specifically, the hypothesis depends upon the *concept of invariance*. Invariance pertains to the *a priori* postulated constancy between descriptors in the two taxonomies.

This discussion, then, will advance the hypothesis that:

(a) The concept of *relatedness* of descriptors can be made numerically precise.

(b) The concept of *relatedness* can serve as a building block for more complex relationships between descriptors.

(c) Some such relationships are postulated as being constant; i.e., these relationships remain invariant in both the external and the ideal taxonomies.

(d) The existence of such constancies forms the basis for selecting rules of reassigning descriptors among documents.

The remainder of this section will attempt to validate this hypothesis and describe the resultant consequences.

4.4.3 *Formalization of the Hypothesis* - Let $d_1, d_2, \ldots, d_n$ and $D_1, D_2, \ldots, D_n$ be descriptors and documents, respectively. For every descriptor there corresponds a class of documents spanned by this descriptor. In set-theoretic notation this concept becomes:

$$[D:\ d(D) = d_i(D)] \qquad (4\text{-}16)$$

which may be read as "the set of all documents such that descriptor d_i applies to the set." To avoid cumbersome notation, the abbreviation $[D(d)]$ will be used to represent the set. The number of documents contained in such a set will be denoted by M. Then $M[D(d_i)]$ stands for the number of documents contained in the set spanned by the descriptor d_i.

In general, every Boolean function of descriptors corresponds a set of documents spanned by these descriptors. Therefore, "the set of all documents that are indexed by B(d)," becomes:

$$[D(B(d))] \tag{4-17}$$

For example,

$$[D(d_1 \wedge (d_2 \vee d_3))] \tag{4-18}$$

is a set of all documents that have as their indices the descriptors d_1 and d_2 or d_3 or both, among others. It is clear that the following relation holds:

$$[D(B(d))] = B[(D(d))] \tag{4-19}$$

This expression signifies that the set of all documents spanned by a Boolean function of descriptors is equivalent to the Boolean function of sets spanned by these descriptors. By analogy, the expression $[d(B(D))]$ represents a set of predicates contained in the set of documents described by the Boolean function B(D).

The relatedness of descriptors or their Boolean functions is defined as the number of documents contained in the intersection of classes spanned by these descriptors or their Boolean functions divided by the number of documents spanned by the union. Formally, this definition becomes:

$$R_d[B_i(d), B_j(d)] = \frac{M[B_i(D(d)) \wedge B_j(D(d))]}{M[B_i(D(d)) \vee B_j(D(d))]} \qquad \text{(Definition 1)} \quad (4\text{-}20)$$

A similar concept of the relatedness of documents or their Boolean functions is defined analogously:

$$R_D[B_i(D), B_j(D)] = \frac{M[B_i(d(D)) \wedge B_j(d(D))]}{M[B_i(d(D)) \vee B_j(d(D))]} \qquad \text{(Definition 2)} \quad (4\text{-}21)$$

It is important to note that throughout this discussion the concepts for descriptors can be analogously applied to documents. The subsequent development, however, will be limited to the relatedness of descriptors.

Since the external taxonomy by hypothesis does not precisely correspond to the ideal taxonomy, the distinct symbol, δ, is introduced to represent the descriptors of the user. These descriptors are only different insofar as they index classes of documents that are not identical with the classes of documents indexed by the descriptors of the external taxonomy. Thus for any descriptor or index i, $[d_i(D)]$ and $[\delta_i(D)]$ are not necessarily identical, even though the descriptors themselves may be the same. The objective of corrective procedures is to adjust the application of descriptors to documents so that the two sets become identical. The corrective procedures may have fulfilled their task if the objective is approximated to the extent that any divergence has a negligible impact upon the user.

4.4.4 The Basis of Corrective Procedures - Assume that all retrieval requests consist of single descriptors. The user formulates his request in terms of a descriptor δ_i related to the ideal taxonomy. The system retrieves all documents spanned by this descriptor, except that this descriptor is d_i in the external taxonomy. The user then decides whether the retrieved collection of documents is satisfactory. The collection may not satisfactorily fulfill the user's requirements for three reasons:

(a) Too many documents were collected.

(b) Too few documents were collected.

(c) Some documents are superfluous and some are missing.

The corrective procedures should select documents more in consonance with user's needs and then effect permanent changes in the application of descriptors to documents.

If the system retrieves too many documents, the system may select a set of descriptors that are most related to the user's descriptor and then remove from the retrieved set those documents spanned by the related descriptors. This method conceals a difficulty. Although a measure for relatedness of two descriptors has been defined, no technique has yet been specified to select clusters of most related descriptors.

If the system retrieves too few documents, a set of descriptors most closely related to the given descriptor is assembled; the set may be limited to a single descriptor. A Boolean function of these descriptors is then constructed, and documents spanned by the Boolean function are retrieved. The factors that determine the nature of the particular Boolean function of descriptors must still be defined.

If some documents are superfluous and some are missing, the problem may be handled as a combination of the specific problems of too many or too few documents. More realistically, however, some problems of this type are *sui generis*, and specific solutions must be developed.

After the originally inadequate set of documents is deleted to the satisfaction of the user, the corrective procedures must effect permanent changes in the extension of some descriptors so that the denotation of the external and ideal descriptors approach equivalence. The problem is to render the sets $[\delta_i(D)]$ and $[d_i(D)]$ extensionally as similar as possible. Several corrective procedures may be used:

(a) To affix the user's descriptor to all the documents and only those documents in the acceptable retrieved set.

(b) To delete or add some descriptors selectively from the set of documents spanned; after the process of deletion or augmentation.

(c) To delete or add some descriptors selectively to the documents that were deleted or complemented from the originally inadequate retrieved set.

(d) To effect other descriptor changes on the documents *not* affected by the processes of complementation or deletion.

The first procedure by itself will not produce the desired transformation until all descriptors have been used in retrieval processes at least once. This prospect is uninviting for any document collection with a large number of descriptors. If such procedure were feasible, there would be no reason not to index the entire collection in the ideal taxonomy, in the first place. In addition, the procedure of complementing the original set of documents need not necessarily lead to the formation of a taxonomy whose extension is identical to the ideal. Rather, the process may only be an approximation; that is, a set obtained after a series of complementations may only approximate the ideal taxonomy.

A closer look at the remaining three procedures and their inherent problems is necessary. Consider a class of documents $[D(d_i)]$ spanned by descriptor d_i. Suppose that the user requests all documents under the descriptor δ_i, a descriptor corresponding to d_i. The class $[D(d_i)]$ is retrieved; it does not fulfill the user's requirements. The complementation procedure results in formation of a new class $[D'(d_i)]$. The corrective procedure should then implement changes pertaining to the distribution of the remaining descriptors among documents. How should these changes be made? Or, to rephrase this question, on what should the inferential processes be based in order to ensure that the ideal taxonomy is approximated?

Assume that there is no relation between the external and the ideal taxonomies. In this case the first stage of the corrective procedure--that is, the complementation of the selected set--must proceed at random. If the taxonomy imposed upon the collection of documents is not correlated with the taxonomy implied by the user, then the relatedness of descriptors to one another will be of no help either in reassigning descriptors or in complementing the original sets.

The possibility of developing corrective procedures depends, therefore, upon some a priori relation between the two taxonomic systems. If such relationships exist, then it must be expressible in terms of the concept of relatedness. The relatedness of descriptors, in one system, must resemble the relatedness in the other. The concept of a relatedness between two taxonomic systems isolates the particular invariance that characterizes the sets of documents designated by certain descriptors.

Formally, an invariance exists if d_iRd_j is true whenever $\delta_iR\delta_j$ is true, where R is a relationship between descriptors. There need not be some universal type of invariance present whenever there is a resemblance between two taxonomic systems. On the contrary, depending upon the nature of the data to be retrieved, the invariance between the ideal and the external taxonomy may differ.

Some examples may clarify the concept of invariance. First, if a set of documents spanned by a descriptor in one system contains another set of documents spanned by another descriptor and if this condition implies the same condition for the corresponding descriptors in the other system, then the invariance might be called <u>nested invariance</u>. Formally:

$$[D(d_j)] \supset [D(d_k)] \rightarrow [D(\delta_j)] \supset [D(\delta_k)] \qquad (4\text{-}22)$$

where → indicates "implies," and ⊃ indicates set inclusion.

In a second example the most closely related descriptors in one system are also most closely related in another. To represent this type of invariance formally, let $(d_i, d_j)^*$ be an ordered pair of descriptors that are related to each other as follows:

$$R_d[(d_i), (d_j)] = \text{Max } R_d[(d_i), (d_k)] \text{ (for all k)} \qquad (4\text{-}23)$$

If then $(d_i, d_j)^* \rightarrow (\delta_i, \delta_j)^*$, the relationship of <u>being most closely related</u> is preserved.

The third example replaces MAX by MIN to obtain an invariance of <u>being the least closely related</u> descriptor. In spite of the formal similarity between the most and least closely related conditions, there is a formidable practical difference. The most closely related condition preserves an invariance between a descriptor and a descriptor; the least closely related condition preserves an invariance between a descriptor and a class of descriptors.

As a fourth example the concept of most closely related descriptors may be applied to chains of descriptors. In such a relationship one descriptor leads to another to form an associative chain. There are many non-equivalent ways of formulating the conditions for the existence of such a chain. One is to let $< d_1, d_2, \ldots, d_n >$ be an associative chain of n^{th} order. Then this chain is defined as:

(a) The set $[d_1, d_2, \ldots, d_n]$ of descriptors comprised in the chain contains each element except the first and the last only once.

(b) The first element appears twice; it is also the last element.

(c) Each element except the first determines its successor by selecting the second most related descriptor. The first descriptor determines its successor by selecting its most related neighbor.

Then, if every associative chain of n^{th} order in one taxonomic system corresponds to a chain in another, a chain invariance of n^{th} order exists. The elements in one chain correspond to the elements in the other, but not necessarily in the same order.

There are a number of additional possible relationships that remain invariant. The problem is to select those that realistically relate to the properties of data structures and their associated indexing systems.

If these invariances exist, rules for reassigning the descriptors may be deduced. The concept of invariance places a strong constraint upon the type of admissible rules that can be formulated. There is also a relation between the invariances and the nature of the convergence and efficiency criteria imposed upon the corrective procedures. The important question is: Given a specific form of invariance and the appropriate rules for complementing sets and for reassigning descriptors, how many queries must elapse before the external taxonomy approximates the ideal? (Approximation in this sense may mean either the probability of obtaining a set that is too small or too large by a specified margin.)

A comparison between one type of invariance and another now becomes possible. These invariances that result in a quick convergence of the corrective procedures are desirable. Conversely, it is possible to investigate the suitability of rules for complementing and reassigning descriptors by keeping a set of invariant relationships constant. All these problems can be investigated mathematically.

4.4.5 Summary - There is an inherent problem in accomodating the descriptors selected for a set of documents by indexing rules to the descriptors used by the user of a system. This problem is related to the extensional difference in the denotation of descriptors or words in an external and an ideal taxonomy. This discussion described methods for developing corrective procedures, which would be applied automatically, to relate the external to the ideal taxonomy. The basis for developing the inferential rules for these procedures is the concept of invariance.

This problem is real, but it is also peripheral. It is more important to develop an adequate indexing concept first; only then does the question of efficiency become important. A significant amount of mathematical formulation remains before the adequate corrective procedures can be implemented.

4.5 MATHEMATICAL MODELS OF FUNCTIONS

4.5.1 General - This section surveys and summarizes some basic concepts of information storage and retrieval and their related mathematical models. These models pertain to particular functions and are thus differentiated from the general system model; in effect, this discussion, which is based upon and derived from Hayes [19, 26], initiates the framework for the formal analysis and development of the transform functions. The elaboration of this framework will be performed during subsequent quarterly periods.

A general theory of information retrieval should encompass at least the following aspects of the problem of storage and retrieval:

(a) Representation of file items.

(b) File organization.

(c) System design and synthesis.

These aspects of a system do not exhaust the elements that should be considered; for example, the measures of relevance presented in the First Quarterly Report also constitute an integral aspect of system design.

A model may be an elegant representation of a trivial problem or a simple representation of a difficult problem. There has been no attempt to evaluate the significance of the following models, since their purpose is to explore the nature of the problems rather than to solve them explicitly and efficiently. Subsequent analysis will be directed to an evaluation of various models in terms of their relation to the system model and their contribution to the solution of the functional problems of information storage and retrieval.

4.5.2 Representation of File Items - The raw material of an information retrieval system consists of documents, requests, and the words or terms used in requests or in referring to or classifying documents. A representation of an element of any of these classes will be called a file item. File items are organized by means of:

(a) Vocabulary.

(b) Syntax.

(c) Coding and format.

Some of the factors of each element of a file item are discussed briefly before a model for item definition is presented.

4.5.2.1 Vocabulary - There are six general types of vocabularies. These types represent a spectrum from unorganized or highly flexible to highly organized or rigidly structured and restrictive. They are listed in order of flexibility, proceeding from the most flexible to the most structured.

(a) Natural language.

(b) Standardized (keywords).

(c) Subject headings.

(d) Semantic factors.

(e) Classifications.

(f) Facet analysis.

The first type is written or conversational language; it permits the use of all the words and phrases in the language, subject only to the rules of grammar and meaning. The second type if restricted to a prescribed set of words; it is discussed by Taube [28] and Jonker [14]. In the third type a restricted set of words is also organized. The fourth type is summarized by Vickery [30]. A more complete development appears in Kent [15]. The fifth type is represented by the well known Dewey decimal and Library of Congress classification systems. The last type is described by Ranganathan [24].

The most common model for describing semantic relations in vocabularies is the lattice. The lattice model is useful primarily because certain lattices can be decomposed into the direct products of two lattices so that vocabulary structures can be exhibited. A theorem to this effect appears in Birkhoff [3].

4.5.2.2 Syntax - A discussion of this area for a sophisticated vocabulary like natural language would be quite discursive and outside the scope of this project. However, for most existing information retrieval systems, a document is represented by a simple conjunction of terms. Correspondingly, a request is represented by a disjunction of conjunctions of terms. The disjunctions indicate separate file items. In a fixed format system such as Uniterm, for example, the syntactical role is a simple one; it is mere presence or absence. However, in some systems the order of terms in a request plays a syntactic role.

4.5.2.3 Coding and Format - Coding and format pertain to the optimal representation of requests and documents. The specific problem is the relationship between the length of the representation code, its effectiveness, and the information that can be retrieved in these terms. The model appropriate for measuring information content in various representations involves information theory, either in a semantic or a classical sense. The classical theory has been will developed, but there has been hardly any development of an information theory based upon semantic concepts.

4.5.2.4 Model for Item Definition - The model in this discussion is geometric; it is not the only possible model. Each document, request, or term to be represented is considered as a physical body that occupies volume and has a mass distribution in a multidimensional space. The volume can be interpreted as the volume of knowledge encompassed by the document; then the mass distribution represents the contribution of a document to each point of knowledge. In an actual retrieval system the body will consist of a discrete set of points in this space. It is assumed that there is a measure of distance and angle in this space so that distances between points and the centers of gravity of sets of points can be computed.

A set of coordinate points is selected and the location of any other point is defined by Barycentric coordinates. In this type of coordinate system, any point is represented as the center of gravity of a distribution of mass at each of the coordinate points. Thus, a point can be located geometrically and assigned a mass.

This general model can be used in several ways to represent file items; one such use is illustrated in Figure 2. A set of key words is chosen as the basic set of coordinate points in the space. Specifically p_1 to p_{12} are the points, each of which represents a key word. P_1, P_2, and P_3 can be interpreted as documents. The key words assigned to P_1 are p_1 through p_5; to P_2, p_6 through p_8; to P_3, p_9 through p_{12}. The documents are located at the center of gravity of their assigned key words. Thus the Barycentric coordinates correspond to the assignment of relative importance of each key word to the document.

P_1, P_2, and P_3 in Figure 2 can also be interpreted as portions of documents. Then the representation of a file item by a set of points is comparable to the representation of a document P by the disjunction of conjunctions of terms. Each conjunction (P_1, P_2, or P_3) represents the center of interest of a major section of the document. If the document treats a relatively restricted topic, as in the case above, one conjunction may be adequate to describe its contents. If it is concerned with several unrelated topics, then several will be required. Each such conjunction corresponds to a single point in the space that was defined. These points are P_1, P_2, and P_3. The information content of any of these points is represented by the set of associated key words. Just as before, each key word defines a point with a given mass, so that the point of interest is the center of gravity of the mass distribution at the key word points. Hence, Barycentric coordinates correspond to the assignment of the relative importance of each key word to a conjunction that represents the information in a portion of a document.

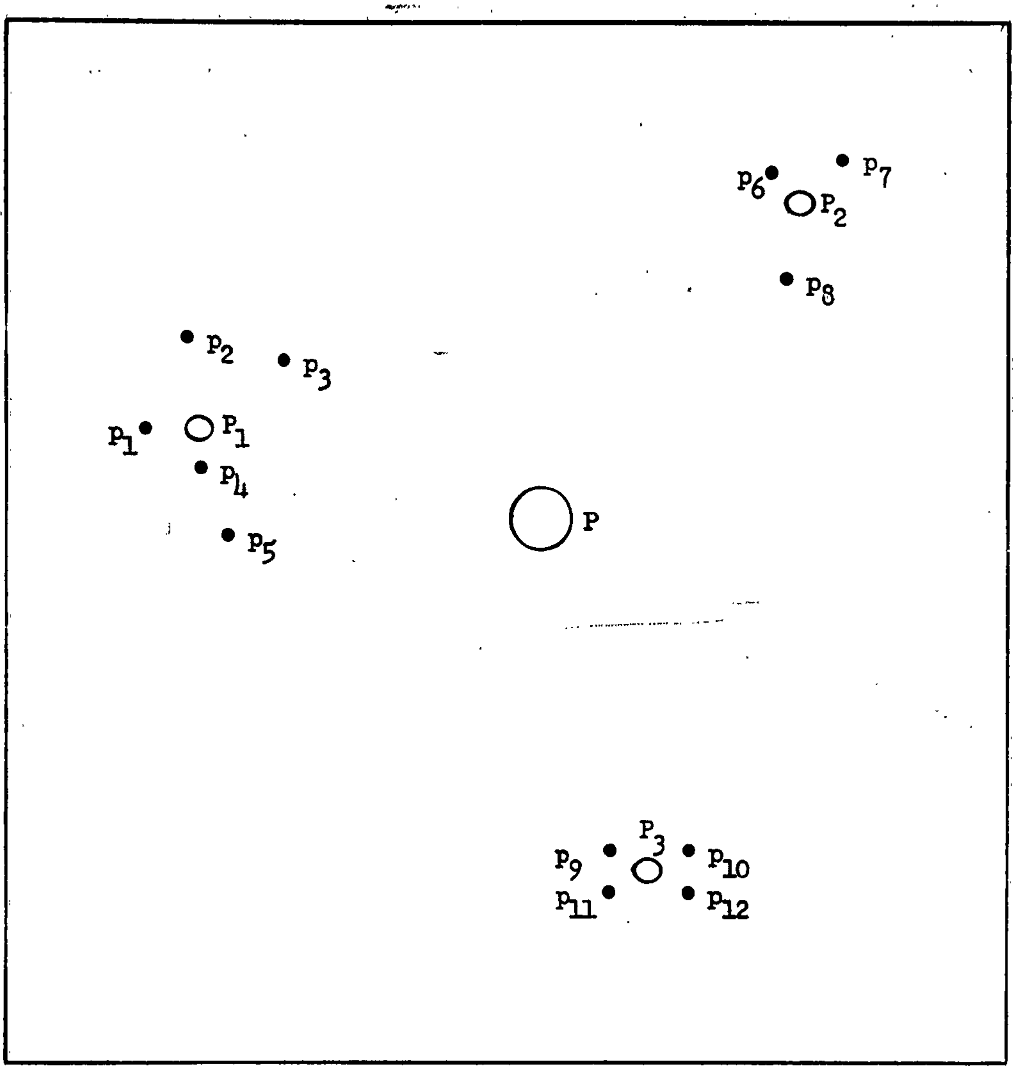

FIGURE 2. Illustration of a Model for Item Definition

In this sense, P is equal to the disjunction $P_1 \vee P_2 \vee P_3$; where P_1, P_2, and P_3 are conjunctions of their associated key words.

4.5.2.5 Relevance - In order to organize and classify terms and documents and to answer requests effectively it is essential to have some measure of the degree of association or relevance of terms or documents. Several such measures were discussed in the First Quarterly Report. Several others could be mentioned: root mean square, Tchebychev sum, minimax, nearness in a Boolean lattice, and the chi-square formula. Most of these measures are either special cases of Barycentric coordinate weightings or are means of order p. A mean of order p for a set of elements x_i is defined as:

$$\sqrt[p]{\Sigma[|x_i|^p]} \qquad (4\text{-}24)$$

Some of these measures can be rejected out-of-hand as counter-intuitive; others would have to be evaluated experimentally.

4.5.3 File Organization - The purpose of file organization is to collect items that are logically related because they are likely to be wanted together whether formally requested or not. A secondary purpose is to improve access time to items that are requested or retrieved frequently. Accordingly, there are four facets of file organization to consider:

(a) Logical organization.

(b) Activity organization.

(c) Physical organization.

(d) Reorganization.

Each facet is analyzed in turn in the following discussion.

4.5.3.1 Logical Organization - The process of coordinate indexing assigns terms to documents. A matrix can be formed with the columns as terms and the rows, documents. An element a_{ij} of the matrix is one or zero depending upon whether the $j^{\underline{th}}$ term is assigned to the $i^{\underline{th}}$ document. This matrix is the document-term matrix. The elements of the document-term matrix can be generalized from a simple YES or NO association to weights that represent the relative importance of the association between a term and a document. The document-term matrix generally assigns many terms to a particular document. Consequently, the retrieval of documents requires the specification of the particular class of pertinent information as a logical conjunction of terms. Boolean algebra or lattice theory is required to specify a particular class of documents.

Although the document-term relationship may be used as a tool for logical organization, the relationships among terms and among documents implicit in the assignment of terms to documents are not fully revealed by the Boolean algebra or lattice structures. For example, the fact that two documents have similar assignments of terms is not apparent from their common assignment to the classes of documents defined by each of the terms. However, this degree of association can be displayed by forming a term-term or document-document matrix. The elements of these matrices would be values of relevance obtained from the document-term matrix by using some previously defined measure of relevance to compare rows or columns.

The objective then is to recover information about the possible groupings of documents or terms from these association matrices. The groups found can be used as classes for defining a generic relationship among terms or as a classification for grouping documents. Several mathematical methods can be used to extract significant factors from an association matrix. They include at least the following: Eigenvalue analysis, factor analysis, powers of the association matrix, and the theory of clumps developed by the Cambridge Language Research Unit in England. These techniques are developed in [1, 4, 5, 8, 12, 21, 26, 29].

Any of these methods produces factors or abstract concepts that are described by relative weightings of the terms or documents from the original set. A relative weighting of the original set of points can be represented by a single point located at the center of gravity of these weights. This point is identical in character to any other point in the space and to any point that might have been chosen to represent a term or a file item. Hence, any set of points related by some degree of association can be grouped into a category labelled with the center of gravity of the set. The abstract terms can by the same methods themselves be grouped into a higher order concept. This technique provides a means for organizing the set of points of the space.

4.5.3.2 Activity Organization - Files can also be organized on the basis of activity; that is, by grouping items according to the likelihood that they will be wanted together. This type of organization can be superimposed upon a logical organization of a file.

The aim of activity organization is to produce a hierarchical arrangement such as nested boxes or levels of grouping. Such an arrangement is illustrated in Figure 3. Each box represents a grouping at some level of abstraction, the level being described by the relative size of the box. The smallest boxes or lowest level contain individual raw file items. If the cover of any box is removed, the interior of the box contains

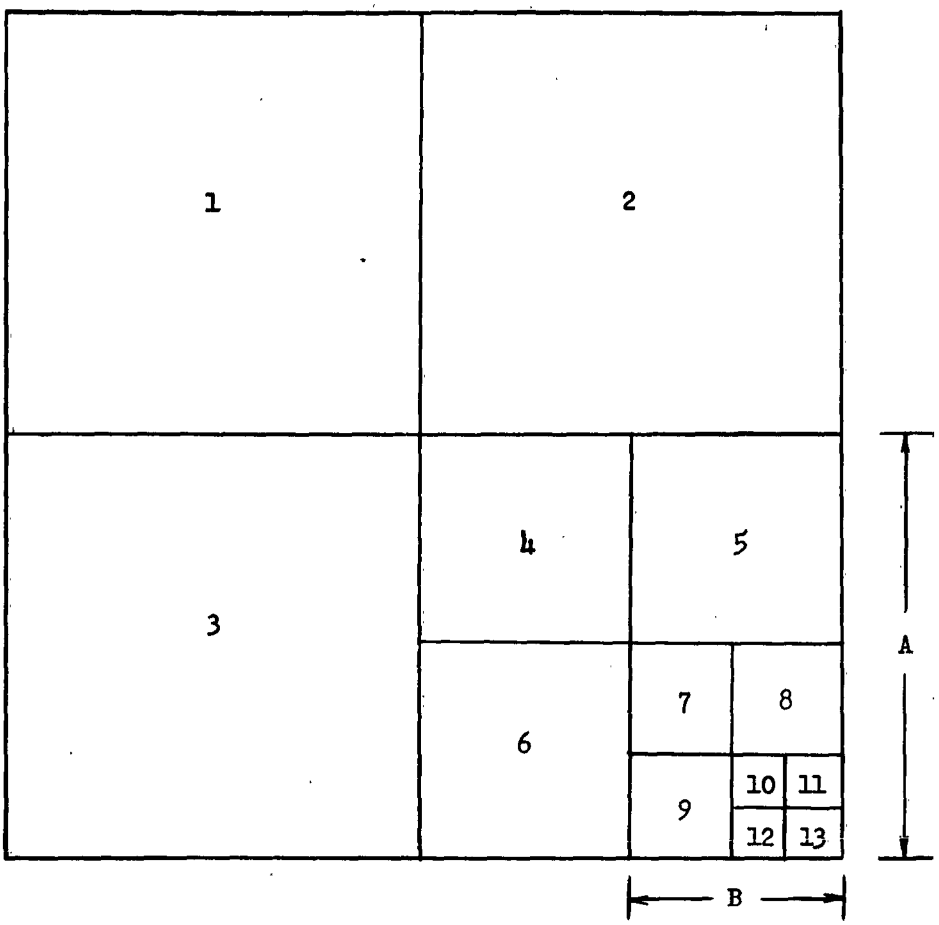

FIGURE 3. Activity Organized File of Nested Boxes

a nest of smaller boxes of the same general character. For example, if the cover of a box labelled 1, 2, or 3 in Figure 3 were removed, it would appear something like the box labelled A; if the covers of boxes 4, 5, and 6 were removed, their contents would look something like box B, which contains boxes 7 to 13.

Each box is labelled by the pattern representing the center of gravity of the patterns contained within it. The actual size of the box at any level is determined from the distribution of the documents in it and their logical relationship. This distribution can be determined from the past concentration of activity, from the value of information contained in the documents, or from a uniform distribution over the space. However, the number of levels (size of boxes) open at any time is dependent upon the *a priori* distribution p(x) of probable activity. The boxes are so designed that the integral of the probability p(x) of each box (independent of its size) is equal for all boxes that are visible at a given time. Then it is equally likely that the answer to any request is in a given visible box independent of its size. For example, if mere entry into the file were the removal of a box cover from the entire file, then the visible box structure might be that of Figure 3. This structure indicates that the probability of finding the answer to a request in box 2 is the same as the probability of finding the answer in box 5. The boxes not visible in Figure 3 represent lower levels of activity in the file.

If a certain box is active, its contents are examined; these contents consist of a set of boxes of equal probable activity. This process is continued until a request is answered satisfactorily by a pattern representing a box at some level, ultimately by a document. Given a measure of the conditional probable activity, given present activity at time t, the boxes are arranged in order according to this measure. The determination of the actual relevance of documents to a situation and the selection of an adequate response involves the matching of a request against the available box patterns; that is, the successive box labels are scanned and matched against the request pattern. The selected box is then opened, and its contents are scanned for a match with the request pattern. This process is continued until the request is answered satisfactorily.

The set of patterns representing the labels of the boxes that are visible at any time is equivalent to an index. The index is scanned, and it indicates where in the space further attention should be directed. The basis for organization by p(x) is that in scanning an index at a certain level, some of the patterns are references to groups of patterns at the next index level, but some are references to lower levels because of the volume of usage of patterns there.

Mathematical expressions, which indicate the number of boxes at each level of a file and the expected number of box covers removed in a search, can be derived in terms of the number of levels of the file and the number of parts in a single partition. The cost of a search is directly related to box size and could be used in addition to relevance as a criterion for selecting boxes whose contents are to be examined.

4.5.3.3 Physical Organization - There is a relation between the logical file organization and the physical organization of a system. The logical file organization can be represented by a tree structure where only the terminal nodes are basic file items; the nodes on other levels represent higher level abstractions. The cost of searching such a tree beginning at the top is a function of the number of levels of the tree, the number of nodes at each level, the number of branches that must be searched, and the access time for each node.

The cost of such a search is a measure of efficiency of the physical organization of the file. If cost is a monotonically increasing function of time, then minimum cost and, therefore, maximum efficiency are achieved in minimum search time. The average search time T can be represented by:

$$T = \sum_{i=1}^{L} \sum_{j=1}^{n_i} P(i,j) \, (t_{aij} + t_{sij}) \qquad (4\text{-}25)$$

where: $P(i,j)$ = probability of selecting j^{th} node, level i

t_{aij} = time to access j^{th} node, level i

t_{sij} = time for selection process j^{th} node, level i

n_i = number of nodes on level i

L = number of file levels

There are two methods for reducing the average search time in such a tree structure. If an estimate of the file activity is available, the order in which the nodes are processed may be revised, allowing a reduction in either or both the access and the process times. This process reflects an activity organization. The second method is to move terminal nodes to a higher level in the tree. Then searches can be terminated without processing all levels of the file (tree).

For activity organization the minimum value of T is obtained when the highest probability is associated with the lowest time (that is, the sum of access and selection times), the next highest probability with the next lowest time, and so on.

For the hierarchical organization, T is minimum when the elements with the highest probability are highest in the tree. Since this type of organization changes the structure of the tree itself, minimum cost C does not necessarily occur at minimum T. Consequently, the criterion for moving the $j^{\underline{th}}$ node from level k to k - 1 is:

$$C_{k-1}T(j,k - 1) < C_kT(j,k) \tag{4-26}$$

In the application of this criterion all nodes are first assigned to the lowest level of the tree and a minimum CT obtained. Moving nodes to the next higher level is then considered in order of their probability. When the criterion is violated, no other moves on that level need be considered because all the remaining nodes on that level have a probability less than or equal to the node that violated the criterion. The nodes on the next higher level are then considered. After the moves from one level to the next are completed, the evaluation begins again at the lowest level of the tree in order to ascertain whether these moves have adversely affected the efficiency of earlier moves. The evaluation moves up the tree until the first new level is processed; then it re-cycles. This procedure is completed when the node with the highest probability violates the criterion or when all levels of the file have been processed.

4.5.3.4 File Reorganization - The usage of information retrieval systems changes with time. Consequently, the distributions upon which an activity organized file are based change with time. On the basis of this and improved knowledge of the value and proper position of documents in the file, a need exists for a procedure that automatically changes the grouping, accessibility, and scanning sequence of file items.

One approach to such a procedure is based upon a multi-level activity organized file with certain logical associations. Suppose stored patterns of bits of fixed word length are divided into three parts called stimulus, response, and index. The stimulus and response sections of the pattern consist of groups of pairs of bits. Each pair of bits corresponds to a particular characteristic of interest. There are four possible values or patterns for a bit pair. Three of these values correspond to values of high, medium, and low for the given characteristic with respect to a particular pattern. Values for some characteristics come from the environment; others are determined from file operations. The bit pair of any characteristic that must be determined by file operation is assigned the fourth possible value, which will be interpreted as a question. The only reason for distinguishing between stimulus and response sections of a pattern is to indicate that the stimulus characteristics are generally prescribed by the environment while the response characteristics are provided by the file. However, this division is not based upon necessity but only upon probability.

The operation of this file may be described with a simple two-level file; the model can be extended without difficulty. The first level stores a limited number of patterns; the second level has the capacity to store an indefinite number of patterns--that is, it will be large enough to handle all patterns not on the first level. In generating patterns the environment prescribes values for certain characteristics and leaves questions for the remainder where values must be supplied by file operations. A partially prescribed pattern of this type is a semi-pattern. The semi-pattern is then matched according to some rule of association with the patterns stored in the first level of the file. This process results in a relative ranking of these patterns in their order of association with the semi-pattern. From the patterns that match the semi-pattern to a degree greater than a specified minimum relevance, those that are most relevant to the semi-pattern are selected. Values for the question bits (characteristics) of the semi-pattern are provided by relative weighting of the corresponding characteristics of the most relevant stored pattern. If none of the stored patterns at the first level have a relevance greater than the prescribed minimum, patterns must be remembered from the next level.

Patterns created from the environmental semi-patterns and file-created answers to questions are stored in the first level. Since the storage capacity of the first level is fixed, it will eventually be exceeded. Therefore, a process must be introduced for forgetting patterns; that is, for transferring patterns to the second level. The procedure is: A quantity is determined by a relative weighting of past relevancy of each first-level pattern and the present relevancy of the pattern to the semi-pattern. In terms of this quantity the least relevant pattern or group of patterns is forgotten. Using the same rule that was used for determining relevance to the environment, the first-level pattern most relevant to the pattern to be forgotten defines the location in which the pattern to be forgotten will be stored. This address is determined from the indexing portion of the relevant pattern. The indexing section of a pattern consists of three addresses: a starting address, the next available address, and the last address assigned in the second level of the file to the relevant pattern. The forgotten pattern is stored in the next available address assigned to the relevant pattern, and the next available address is updated.

When the patterns stored at the first level do not match a semi-pattern to the specified minimum degree of relevance, patterns must be remembered or recalled from the second level by means of the indices of the most relevant patterns, even though they are below the acceptable minimum. The index section of the most relevant pattern at the first level thus provides a mechanism for obtaining a pattern from the second level, bringing it to the first level, and examining it for relevancy. This process is continued until sufficiently relevant patterns are found or until no further index data is available. If neither of these conditions occurs after a reasonable prescribed time, the process can be stopped arbitrarily; alternatively, the process can be stopped whenever a new semi-pattern is accepted.

4.5.4 System Design and Synthesis - Detailed consideration of system design and synthesis should be postponed until the other areas have been developed to a greater extent. The other areas are not system oriented, while this one is. It therefore constitutes the last phase in the development of a theory of information retrieval. A convenient subdivision of this phase is:

(a) Organization of processes.

(b) Organization of equipment and personnel.

(c) Evaluation of system efficiency.

For the sake of completeness, this area will be discussed briefly.

4.5.4.1 Organization of Processes - The organization of processes is sometimes called the logical design of a system. The end product is usually a set of flow charts. These charts would show the sequence of functions to be performed, decisions and alternatives, points of interrelation and feedback, and the inputs and outputs for each function. There is as yet no adequate mathematical method for isolating system functions and completing the logical design. The resultant flow charts, however, do serve as a sort of schematic graphical model of the system design.

4.5.4.2 Organization of Equipment and Personnel - The objective of this area is to allocate tasks or assign functions to equipment and personnel. Criteria for these allocations are the flexibility, speed, and accuracy requirements of the various functions and subfunctions comprising the system. To date the allocation of functions to men and machines has been an art largely constrained by the rigidity of computer techniques for associating, classifying, storing, and retrieving data. In other words, all those functions that could not be automated with the required degree of flexibility have been allocated to personnel. Improvements in this function, therefore, will not depend upon mathematicizing the process but upon developing better mathematical models in the areas of file item representation, file organization, and evaluation of system efficiency, and related problem areas.

4.5.4.3 Evaluation of System Efficiency - Adequate criteria for measuring the value of an information system have not yet been developed. Therefore, models of system efficiency must be viewed as aids to design, which may confirm intuitive judgments, but not as adequate tools in themselves to make design decisions.

An information system is a collection of components that in concert perform a set of operations to accomplish a specific purpose. The system is represented by a matrix A of efficiency values. The rows of this matrix correspond to the individual operations. The element e_{ij} is the efficiency with which the $i^{\underline{th}}$ component performs on the $j^{\underline{th}}$ operation. The component efficiencies could be defined by some parameter such as the product of cost in dollars per unit of time and the operational time divided by the number of bits processed; that is, the efficiency has units of dollars per bit. A volume vector v can be defined whose components are the volumes or traffic loads for each operation. The product of the efficiency matrix and volume vector is defined as the required cost vector C whose components are the costs required to perform the given volume of the set of operations at the defined efficiencies. There are practical problems in determining the various parameters, but these will be ignored in illustrating the model. Using a rms measure of efficiency E yields:

$$\frac{1}{E} = \sqrt{\frac{C^* C}{v^* v}} = \sqrt{\frac{v^* A^* A v}{v^* v}} \qquad (4\text{-}27)$$

where A^* indicates the transpose of the matrix A. The quantity under the radical on the right is the Rayleigh quotient for the matrix $A^* A$. Efficiency can now be maximized by the methods of Eigenvalue analysis. A generalization of the classical Eigenvalue theory is required to handle a non-square matrix A, directly. This mathematical generalization is available in Hestenes [13].

4.5.5 Summary - This section discussed some mathematical models and their purposes as related to specific problem areas in information retrieval. These models are related in the following coherent summary:

(a) Vocabulary

(1) Objective - Description of semantic relations
(2) Data Source - Vocabularies
(3) Model - Lattice

(b) Coding and Format

(1) Objective - Measurement of information content
(2) Data Source - Document abstract size
(3) Model - Information theory

(c) Logical Organization of Files

(1) Objective - Measurement of relevancy and categorization in terms of it
(2) Data Source - Document-term matrix
(3) Model - Matrix algebra

(d) Activity Organization of Files

(1) Objective - Measurement and optimization of responsiveness
(2) Data Source - Activity distributions
(3) Model - Nested box structures

(e) Physical Organization of Files

(1) Objective - Optimization of physical organization of files
(2) Data Source - Facility costs and operating rates, personnel costs and operating rates, sequence of operations, and activity distributions
(3) Model - Average cost of a search

(f) Reorganization of Files

(1) Objective - Definition of programmable processes for file reorganization
(2) Data Source - Statistics of environment
(3) Model - Multi-level index-connected file

(g) System Efficiency

(1) Objective - Measurement and optimization of system efficiency
(2) Data Source - Component-operation performance analysis resulting in the component-operation efficiency matrix
(3) Model - Matrix algebra

It should be emphasized that these models are not necessarily the best nor the only models that can be developed to solve any particular problem.

4.6 REFERENCES

(1) Baker, F. B., "Information Retrieval Based on Latent Class Analysis," Journal of the ACM, Vol. 9, No. 4; October 1962: pp 512-521.

(2) Baxendale, P. B., "Machine-Made Index fpr Technical Literature--An Experiment," IBM Journal of Research and Development, Vol. 2, No. 4; October 1958: pp 354-361.

(3) Birkhoff, G., Lattice Theory; The American Mathematical Society Colloquium Publications, 1948.

(4) Bodewig, E., Matrix Calculus; North-Holland Publishing Co., 1959.

(5) Borko, H., The Construction of an Empirically Based Mathematically Derived Classification System (AD 267901); System Development Corporation (Report SP-585), October 1961.

(6) Borko, H., and Bernick, M. D., Automatic Document Classification; System Development Corporation, (Technical Memorandum TM-771), November 1962.

(7) Edmundson, H. P., and Wyllys, R. E., "Automatic Abstracting and Indexing: Survey and Recommendations," Communications of the Association for Computing Machinery, Vol. 4, No. 5; May 1961: pp 226-234.

(8) Faddeeva, V. N., Computational Methods of Linear Algebra; Dover Publications, Inc., 1959.

(9) Gray, H. J., et al, Information Retrieval and the Design of More Intelligent Machines; Final Report No. AD59URI to the U. S. Signal Corps, July 1959.

(10) Gray, H. J., et al, The Multi-List System; Report to the Office of Naval Research, Information Systems Branch, under Contract NOnr551(40), November 1961.

(11) Green, B. F., Wolf, A. K., Chomsky, Carol, and Laughery, K., "Baseball: An Automatic Question-Answerer," Proceedings WJCC; IRE, Los Angeles, May 1961.

(12) Harmon, H. H., Modern Factor Analysis; University of Chicago Press, 1960.

(13) Hestenes, M. R., "Inversion of Matrices by Biorthogonalization and Related Results," Journal of Society for Industrial and Applied Mathematics; March 1958.

(14) Jonker, F., The Descriptive Continuum - A Generalized Theory of Indexing; Air Force Office of Scienctific Research, June 1957.

(15) Kent, Allen, and Perry, J. W., Technical Notes (series), Center for Documentation and Commmication Research, School of Library Science, Western Reserve University.

(16) Luhn, H. P., "A Statistical Approach to Mechanized Encoding and Searching of Literary Information," IBM Journal of Research and Development, Vol. 1, No. 4; October 1957: pp 309-317.

(17) Maron, M. E., "Automatic Indexing: An Experimental Inquiry," Journal of the Association for Computing Machinery, Vol. 8, No. 3; July 1961: pp 404-417.

(18) Maron, M. E., and Kuhns, J. L., "On Relevance, Probabilistic Indexing and Information Retrieval," Journal of the Association for Computing Machinery, Vol. 7, No. 2; July 1960: pp 216-244.

(19) Mathematical Models for Information System Design and a Calculus of Operations, Final Report; Advanced Information Systems Co., Air Force Contract AF 30(602)-2111, 1961.

(20) Oswald, V. A., Jr., et al, Automatic Indexing and Abstracting of the Contents of Documents, (RADC-TR-59-208); Prepared for the Rome Air Development Center, Air Research and Development Command, United States Air Force, 31 October 1959: pp 5-34, 59-133.

(21) Parker-Rhodes, A. F., and Needham, R. M., The Theory of Clumps; Cambridge Language Research Unit, Cambridge, England, February 1960.

(22) Perry, J. W., Kent, A., and Berry, M. M., Machine Literature Searching; New York, 1956.

(23) Personal communications and informal briefing.

(24) Ranganathan, S. R., Classifying, Indexing, Coding; Western Reserve University, September 1959.

Classification and Retrieval - Problems for Pursuit; Western Reserve University, September 1959.

Natural, Classificatory, and Machine Languages; Western Reserve University, September 1959.

(25) Rath, G. J., Resnick, A., and Savage, T. R., Comparisons of Four Types of Lexical Indicators of Contents, (Research Report RC-187); IBM Research Center, Yorktown Heights, New York, 14 August 1959.

(26) Report on the Organization of Large Files with Self-Organizing Capability; Advanced Information Systems Co., National Science Foundation Contract C 162, 1961.

(27) Stiles, H. E., The Association Factor in Information Retrieval, Journal of the Association for Computing Machinery, Vol. 8, No. 2, April 1961: pp 271-279.

(28) Taube, M., et al, Studies in Coordinate Indexing; Documentation Incorporated, 1953-57.

(29) Thurstone, L. L., Multiple Factor Analysis; University of Chicago Press, 1947.

(30) Vickery, B. C., Journal of the American Documentation Institute, Vol. X, 1959: pp 234-241.

(31) Vickery, B. C., On Retrieval Systems Theory; Butterworths, London, 1961.

(32) Watanabe, S., Inference and Information; John Wiley & Sons, New York, (to be published).

5. CONCLUSIONS

Four aspects of the research orientation were described in establishing the frame of reference for this project: system-procedure, real-hypothetical, hardware-software, reduction-manipulation. A theoretical--procedural, hypothetical, software, manipulative--approach has been adopted. A preliminary generalized model has been formulated as a basis for analyzing detailed aspects of the problem. Several procedural areas have been analyzed in varying degrees of formalization. The interrelationships among the functional characteristics of the preliminary model as well as their relation to the entire problem are being investigated. There remains an extensive task of formalizing these areas into an integrated whole in order to fulfill the objectives of the program.

6. PLANS FOR THE NEXT QUARTER

Activities during the next quarter will proceed with the over-all goal of developing a theory of information retrieval for use as a tool in the design of information retrieval systems. Work will include at least the following three aspects of the development of such a theory.

(a) A statement of the necessary or desirable features of a theory of information retrieval together with a breakdown of the essential functional elements of information retrieval and their interrelationships.

(b) Continue development of an information retrieval model based on Item (a) and the preliminary model. This work will include utilizing and relating results of Item (c).

(c) Continue work on functional elements of the model and techniques that are applicable to the effective performance of these essential functions (e.g., measures of relevance as applied to descriptor assignment).

These three aspects of the work are actually levels of detail. The first provides a general statement of the objectives of the research, defines essential areas of effort, and provides guidelines and definitions for use in the development of the theory. The second level of effort develops and defines the essential features of the theory to the point where a representative model is meaningful. It will isolate independent functions and establish relations between functions that are not independent. The third level develops detailed techniques, procedures, and methodology useful for the design of an effective information retrieval system.

During the next quarter each aspect of activity will also be oriented to the definition, development, and exposition of specific tasks within this general methodological framework.

7. IDENTIFICATION OF PERSONNEL

7.1 PERSONNEL ASSIGNMENTS

The following personnel were assigned to the project during the period covered by this report:

Name	Title	Man-Hours
Jacques Harlow	Manager	60
Quentin A. Darmstadt	Research Specialist	260
George Greenberg	Senior Specialist	300
Alfred Trachtenberg	Senior Program Analyst	425

The man-hours applied to the project during this period deviated slightly from the schedule because of conferences, holidays, and vacations--all of which were heavily concentrated during this reporting period.

7.2 BACKGROUND OF PERSONNEL

The backgrounds of the personnel assigned to the project were described in the First Quarterly Report. No new personnel were assigned to the project.

8. APPENDICES

8.1 APPENDIX A - Maxima and Minima of the Measures

The behavior of the measures of goodness and the various entropy functions will now be examined. Maxima and minima in terms of the p_j and p_{ij} are summarized in Tables 1 and 2.

For these tables, it is assumed that A is chosen such that $A = \frac{1}{p_e}$ where p_e is the smallest p_j; that is, $p_e \leq p_j$ for all j. For the functions of Table 1--H, H_i, H_A and S_i--the pertinent values are the maximum and minimum values in terms of a given p_e and the absolute maximum and minimum values of each function.

For H and H_i, maxima are reached when the probabilities are equal or, for a particular p_e, when the other p_j are equal; minima are reached when one probability becomes a maximum and the rest are minima.

While H_A does not reach an absolute maximum when H does, since it was assumed that $A = \frac{1}{p_e}$, it does reach a maximum together with H for a particular p_e. Then:

$$H_A = - \sum_j p_j \log p_j + \log A = - \sum_j p_j \log p_j - \log p_e$$

$$= - \sum_{j \neq e} p_j \log p_j - (1 + p_e) \log p_e \qquad (8\text{-}1)$$

Therefore, H_A becomes a maximum for a particular p_e when $p_j = \frac{1 - p_e}{k - 1}$ for $j \neq e$. Then:

$$H_{A\max} = (1 - p_e) \log \left(\frac{k - 1}{1 - p_e}\right) - (1 + p_e) \log p_e \qquad (8\text{-}2)$$

The largest $H_{A\max}$ occurs when $p_e = 1/N$. Then:

TABLE 1. Maxima and Minima of Entropy Functions

Function	Max occurs at:	Max is:
H	max when $p_j = \frac{1-p_e}{k-1}$ for $j \neq e$ $p_e \leq p_j$ for all j	$H_{max} = (1 - p_e) \log (\frac{k-1}{1-p_e}) - p_e \log p_e$
	abs max when all p_j are equal $p_j = p_e = \frac{1}{k}$	$H_{absmax} = \log k$
H_i	$p_{ij} = \frac{1}{k}$ for all j	$H_{imax} = \log k$
H_A	max when $p_j = \frac{1-p_e}{k-1}$ for $j \neq e$ $p_e \leq p_j$ for all j	$H_{Amax} = (1 - p_e) \log (\frac{k-1}{1-p_e}) - (1 + p_e) \log p_e$
	abs max when $p_e = \frac{1}{N}$	$H_{Aabsmax} = (1 + \frac{1}{N}) \log N + (1 - \frac{1}{N}) \log (\frac{k-1}{1-\frac{1}{N}})$
S_i	max when $p_{ij} = p_j$ for all j	$S_{imax} = -\log p_e$
	abs max when $p_e = \frac{1}{N}$	$S_{iabsmax} = \log N$

TABLE 1 (Continued). Maxima and Minima of Entropy Functions

Function	Min occurs at:	Min is:
H	min when $p_t = 1 - (k - 1)p_e$, $p_j = p_e \ (j \neq t)$ $p_e \leq p_j$ for all j	$H_{min} = -(k - 1)p_e \log p_e - [1 - (k - 1)p_e] \log[1 - (k - 1)p_e]$
	abs min when $p_j = p_e = \frac{1}{N} \ (j \neq t)$ $p_t = 1 - \frac{k-1}{N}$	$H_{absmin} = \frac{k-1}{N} \log N - (1 - \frac{k-1}{N}) \log (1 - \frac{k-1}{N})$
H_i	$p_{ie} = 1$, $p_{ij} = 0 \ (j \neq e)$	$H_{imin} = 0$
H_A	min when $p_t = 1 - (k - 1)p_e$ $p_j = p_e \ (j \neq t)$ $p_e \leq p_j$ (for all j)	$H_{Amin} = -[1 + (k - 1)p_e] \log p_e - [1 - (k - 1)p_e] \log[1 - (k - 1)p_e]$
	abs min when $p_e = p_j = \frac{1}{k}$	$H_{Aabsmin} = 2 \log k$
S_i	$p_{ie} = 1$, $p_{ij} = 0$ for $j \neq e$ $p_e \leq p_j$ for all j	$S_{imin} = 0$

$$H_{Aabsmax} = (1 + \frac{1}{N}) \log N + (1 - \frac{1}{N}) \log (\frac{k - 1}{1 - \frac{1}{N}}) \qquad (8\text{-}3)$$

H_A becomes a minimum for a particular p_e when H does; that is, when the maximum p_j, $p_t = 1 - (k - 1) p_e$, and $p_j = p_e$ for $j \neq t$, where $p_e \leq p_j$ for all j. Then:

$$H_{Amin} = -[1 - (k - 1) p_e] \log[1 - (k - 1) p_e] -[1 + (k - 1) p_e] \log p_e \qquad (8\text{-}4)$$

The smallest H_{Amin} occurs when $p_e = 1/k$. Then:

$$H_{Aabsmin} = 2 \log k \qquad (8\text{-}5)$$

S_i becomes a maximum when $p_{ij} = p_j$ for all j. This maximum can be derived by using Gibbs' theorem, as in Watanabe [32]:

$$S_{imax} = \log A = - \log p_e \qquad (8\text{-}6)$$

The largest S_{imax} occurs when $p_e = 1/N$.

$$S_{iabsmax} = \log N \qquad (8\text{-}7)$$

S_i becomes a minimum when p_{ij} becomes one for the particular j for which p_j is smallest. Then:

$$S_{imin} = - \log \frac{1}{Ap_e} \qquad (8\text{-}8)$$

but $$A = 1/p_e \qquad (8\text{-}9)$$

so $$S_{imin} = 0 \qquad (8\text{-}10)$$

For the functions of Table 2--M_1, M_2, M_3, and M_4--there are three

TABLE 2. Maxima and Minima of Measures of Goodness

Function	Max occurs at:	Max is:
$M_1 = H = H_i$	max for p_j when: $H_i = H_{imin}$	$M_{1maxj} = H$
	max when: $H = H_{max}$, $H_i = H_{imin}$	$M_{1max} = H_{max}$
	abs max when: $H = H_{absmax}$, $H_i = H_{imin}$	$M_{1absmax} = H_{absmax} = \log k$
$M_2 = H - S_i$	max for p_j when: $S_i = S_{imin}$	$M_{2maxj} = H$
	max when: $H = H_{max}$, $S_i = S_{imin}$	$M_{2max} = H_{max}$
	abs max when: $H = H_{absmax}$, $S_i = S_{imin}$	$M_{2absmax} = H_{absmax} = \log k$
$M_3 = H_A - S_i$	max for p_j when: $S_i = S_{imin}$	$M_{3maxj} = H_A$
	max when: $H_A = H_{Amax}$, $S_i = S_{imin}$	$M_{3max} = H_{Amax}$
	abs max when: $H_A = H_{Aabsmax}$, $S_i = S_{imin}$	$M_{3absmax} = (1 + \frac{1}{N}) \log N + (1 - \frac{1}{N}) \log(\frac{k-1}{1-\frac{1}{N}})$
$M_4 = \log A - S_i$	max for p_j when: $S_i = S_{imin}$	$M_{4maxj} = \log A = -\log p_e$
	max when: $S_i = S_{imin}$	$M_{4max} = \log A = -\log p_e$
	abs max when: $S_i = S_{imin}$, $p_e = \frac{1}{N}$	$M_{4absmax} = \log N$

TABLE 2 (Continued). Maxima and Minima of Measures of Goodness

Function	Min occurs at:	Min is:
$M_1 = H - H_i$	min for p_j when: $H_i = H_{imax}$	$M_{1minj} = H - H_{imax} = H - \log k$
	min when: $H = H_{min}, H_i = H_{imax}$	$M_{1min} = H_{min} - H_{imax} = H_{min} - \log k$
	abs min when: $H = H_{absmin}, H_i = H_{imax}$	$M_{1absmin} = -\log k + \frac{k-1}{N} \log N$ $-(1 - \frac{k-1}{N}) \log(1 - \frac{k-1}{N})$
$M_2 = H - S_i$	min for p_j when: $S_i = S_{imax}$	$M_{2minj} = H - S_{imax} = H + \log p_e$
	min when: $H = H_{min}, S_i = S_{imax}$	$M_{2min} = H_{min} - S_{imax} = H_{min} + \log p_e$
	abs min when: $H = H_{absmin}, S_i = S_{iabsmax}$	$M_{2\,bsmin} = (\frac{k-1}{N} - 1) \log N$ $-(1 - \frac{k-1}{N}) \log(1 - \frac{k-1}{N})$
$M_3 = H_A - S_i$	min for p_j when: $S_i = S_{imax}$	$M_{3minj} = H$
	min when: $H = H_{min}, S_i = S_{imax}$	$M_{3min} = H_{min}$
	abs min when: $H = H_{absmin}, S_i = S_{imax}$	$M_{3absmin} = \frac{k-1}{N} \log N$ $-(1 - \frac{k-1}{N}) \log(1 - \frac{k-1}{N})$
$M_4 = \log A - S_i$	min for p_j when: $S_i = S_{imax}$	$M_{4minj} = 0$
	min when: $S_i = S_{imax}$	$M_{4min} = 0$
	abs min when: $S_i = S_{imax}$	$M_{4absmin} = 0$

maximum and minimum values: the maxima and minima for a given p_j distribution; the maxima and minima when only p_e is given; and the absolute maxima and minima. To keep the notation consistent with that of Table 1, these maxima and minima will be indicated as follows:

M_{1maxj}, M_{2maxj}, etc.

are the maxima for a given p_j distribution. Similarly,

M_{1minj}, M_{2minj}, etc.

are the minima for a given p_j distribution.

M_{1max}, M_{2max}, M_{1min}, M_{2min}, etc., are the maxima and minima when only p_e is given, and $M_{1absmax}$, $M_{2absmax}$, $M_{1absmin}$, $M_{2absmin}$, etc., are the absolute maxima and minima.

$M_1 = H - H_i$ is maximized for a particular p_j distribution when H_i is a minimum ($H_{imin} = 0$). Then M_{1maxj} is simply the a priori entropy H. M_{1max}, which is M_1 maximized for a particular p_e, is simply the a priori entropy maximized, H_{max}. $M_{1absmax}$ is the absolute maximum of the a priori entropy.

Similarly the minima of M_1 are obtained when H_i is set equal to H_{imax} ($H_{imax} = \log k$) by minimizing the a priori entropy.

$M_2 = H - S_i$ is maximized when S_i is a minimum ($S_{imin} = 0$); the maxima are simply the maxima of the a priori entropy. M_2 is minimized when $S_i = S_{imax} = - \log p_e$; $M_{2min} = H_{min} - S_{imax}$ when $H = H_{min}$ in addition. $M_{2absmin}$ occurs when $H = H_{absmin}$. $M_3 = H_A - S_i$ is maximized when $S_i = S_{imin}$; the maxima are H_A, H_{Amax}, and $H_{Aabsmax}$, respectively. The minima of M_3 are not as obvious, for the conditions of maximizing S_i and minimizing H_A can be contradictory. It is best to analyze the minima of M_3 as follows:

$$M_3 = H_A - S_i = -\sum_j p_j \log p_j + \log A + \sum_j p_{ij} \log \frac{p_{ij}}{Ap_j}$$

$$= -\sum_j p_j \log p_j + \sum_j p_{ij} \log \frac{p_{ij}}{p_j} \qquad (8\text{-}11)$$

For a particular p_j distribution, M_{3minj} occurs when $p_{ij} = p_j$ for all j. Therefore:

$$M_{3minj} = -\sum_j p_j \log p_j = H \qquad (8\text{-}12)$$

Then for a particular p_e:

$$M_{3min} = H_{min}, \qquad (8\text{-}13)$$

and the absolute minimum is simply:

$$M_{3absmin} = H_{absmin}. \qquad (8\text{-}14)$$

M_4 is the simplest measure of them all, reaching a maximum when S_i is minimum, and a minimum when S_i is maximum.

$$M_4 = \log A - S_i = +\sum_j p_{ij} \log \frac{p_{ij}}{p_j} \qquad (8\text{-}15)$$

That this measure is always greater than or equal to zero can be shown by applying Gibbs' theorem:

$$M_4 = \sum_j p_{ij} \log p_{ij} - \sum_j p_{ij} \log p_j \qquad (8\text{-}16)$$

But: $\sum_j p_{ij} \log p_{ij} - \sum_j p_{ij} \log p_j \geq 0$ (Gibbs' theorem) (8-17)

Therefore, $M_4 \geq 0$. (8-18)

The maximum of M_4 is:

$$M_{4\max j} = M_{4\max} = \log A \qquad (8\text{-}19)$$

The absolute maximum occurs when:

$$p_e = \frac{1}{N}; \quad \text{then } A = N \text{ and } M_{4\text{absmax}} = \log N \qquad (8\text{-}20)$$

DISTRIBUTION LIST

Recipient	Copies
OASD (R&E) Rm 3E1065 Attention: Technical Library The Pentagon Washington 25, D. C.	1
Chief of Research and Development OCS, Department of the Army Washington 25, D. C.	1
Commanding General U. S. Army Materiel Command Attention: R&D Directorate, Res Div, Elect Br. Washington 25, D. C.	1
Commanding General U. S. Army Electronics Command Attention: AMSEL-AD Fort Monmouth, New Jersey	3
Commander, Armed Services Technical Information Agency Attention: TIPOR Arlington Hall Station Arlington 12, Virginia (Reports)	10
Commanding General USA Combat Developments Command Attention: CDCMR-E Fort Belvoir, Virginia	1
Commanding Officer USA Communication and Electronics Combat Development Agency Fort Huachuca, Arizona	1
Commanding General U. S. Army Electronics Research and Development Activity Attention: Technical Library Fort Huachuca, Arizona	1

Recipient	Copies
Chief, U. S. Army Security Agency Arlington Hall Station Arlington 12, Virginia	2
Deputy President U. S. Army Security Agency Board Arlington Hall Station Arlington 12, Virginia	1
Director, U. S. Naval Research Laboratory Attention: Code 2027 Washington 25, D. C.	1
Commanding Officer and Director U. S. Navy Electronics Laboratory San Diego 52, California	1
Aeronautical Systems Division Attention: ASAPRL Wright-Patterson Air Force Base, Ohio	1
Air Force Cambridge Research Laboratories Attention: CRZC L. G. Hanscom Field Bedford, Massachusetts	1
Air Force Cambridge Research Laboratories Attention: CRXL-R L. G. Hanscom Field Bedford, Massachusetts	1
Headquarters, Electronic Systems Division Attention: ESAT L. G. Hanscom Field Bedford, Massachusetts	1
Rome Air Development Center Attention: RAALD Griffiss Air Force Base, New York	1

Recipient	Copies
AFSC Scientific/Technical Liaison Office U. S. Naval Air Development Center Johnsville, Pennsylvania	1
Commanding Officer U. S. Army Electronics Materiel Support Agency Attention: SELMS-ADJ Fort Monmouth, New Jersey	1
Director, Fort Monmouth, Office USA Communication and Electronics Combat Development Agency Fort Monmouth, New Jersey	1
Corps of Engineers Liaison Office U. S. Army Electronics Research & Development Laboratory Fort Monmouth, New Jersey	1
Marine Corps Liaison Office U. S. Army Electronics Research & Development Laboratory Fort Monmouth, New Jersey	1
AFSC Scientific/Technical Liaison Office U. S. Army Electronics Research & Development Laboratory Fort Monmouth, New Jersey	1
Commanding Officer U. S. Army Electronics Research & Development Laboratory Attention: Logistics Division Fort Monmouth, New Jersey Attention: Anthony V. Campi	9
Commanding Officer U. S. Army Electronics Research & Development Laboratory Attention: Director of Research/Engineering Fort Monmouth, New Jersey	2
Commanding Officer U. S. Army Electronics Research & Development Laboratory Attention: Technical Documents Center Fort Monmouth, New Jersey	2

Recipient	Copies
Commanding Officer U. S. Army Electronics Research & Development Laboratory Attention: SELRA/NPE Fort Monmouth, New Jersey	2
Commanding Officer U. S. Army Electronics Research & Development Laboratory Attention: Technical Information Division Fort Monmouth, New Jersey	3
Commanding Officer U. S. Army Electronics Research & Development Laboratory Attention: Exploratory Research Dr. Reilly Fort Monmouth, New Jersey	2
Commanding Officer U. S. Army Electronics Research & Development Laboratory Attention: Engineering Sciences Department Mr. Hennessy Fort Monmouth, New Jersey	2
Commanding Officer U. S. Army Electronics Research & Development Laboratory Attention: Exploratory Research Jack Benson Fort Monmouth, New Jersey	3